

DOCUMENT RESUME

ED 269 459

TM 860 313

AUTHOR O'Neill, Kathleen A.
TITLE The Effect of Stylistic Changes on Item Performance.
PUB DATE Apr 86
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Abbreviations; Analysis of Variance; Drug Education; Graduate Medical Students; Higher Education; *Item Analysis; *Language Skills; Multiple Choice Tests; Multivariate Analysis; Pharmaceutical Education; Symbols (Literary); Test Construction; *Test Format; *Testing Problems; *Test Items
IDENTIFIERS *Bonferroni Procedure; *Confounding Variables

ABSTRACT

When test questions are not intended to measure language skills, it is important to know if language is an extraneous characteristic that affects item performance. This study investigates whether certain stylistic changes in the way items are presented affect item performance on examinations for a health profession. The subjects were medical residents taking a multiple-choice in-training examination. Twenty items were selected from the pretest section of the examination to cover three stylistic manipulations: (1) abbreviations; (2) symbols; and (3) generic and proprietary drug names. Every item was tested in two versions (unedited, edited) and these item pairs involved the same stylistic change for each pair. A Bonferroni test of significance was applied to the differences across forms for each item pair. A two-way multivariate analysis of variance yielded essentially the same results as the two-way analysis of variance. No significant differences in performance were noted for individual item pairs or for type of stylistic manipulation. However, a test form effect was noted on one of the two replications. Test developers need to be aware that even if there is no difference in statistics on an item level, differences might exist on a test form level. (Author/PN)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED269459

The Effect of Stylistic Changes on Item Performance

Kathleen A. O'Neill

The American College Testing Program

Paper Presented at the Annual Meeting
of the AERA and NCME,
San Francisco, CA, 1986

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

K. A. O'Neill

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Abstract

When test questions are not intended to measure language skills, it is important to know if language is an extraneous characteristic that affects item performance. Twenty items were selected from the pretest section of a medical in-training examination to cover three stylistic manipulations: 1) abbreviations, 2) symbols, and 3) generic and proprietary drug names. Every item was tested in two versions (unedited, edited) and these item pairs involved the same stylistic change for each pair. No significant differences in performance were noted for individual item pairs or for type of stylistic manipulation. However, a test form effect was noted on one of the two replications. Test developers thus need to be aware that even if there is no difference in statistics on an item level, differences might exist on a test form level.

The Effect of Stylistic Changes on Item Performance

Introduction

Researchers have long been interested in whether test-item language influences examinees' responses to items. When test questions are not intended to measure language skills, it is important to know if language is an extraneous item characteristic that affects item performance. To the extent that minor stylistic changes can affect the way examinees respond, the validity of the test is affected. Many studies on language have investigated such item characteristics as grammatical cues (Board and Whitney, 1972; Plake and Huntley, 1984) or response option length (McMorris et al, 1972). In fact, a recent survey of item-writing research notes that studies of option length, "window dressing," and clueing are among the most popular topics (Haladyna and Downing, 1985).

A different facet of research on language is the complexity of the language itself. Studies of this kind have manipulated vocabulary level and syntactic complexity within the context of general knowledge items. Research in this area has noted mixed results. Some studies have concluded that language difficulty does affect item performance in a minor way (Millman, 1978; Green, 1984); others have found no difference (Bornstein and Chamberlain, 1978; Jerman and Mirman, 1971).

The present study investigates whether certain stylistic changes in the way items are presented affect item performance on examinations for a specific health profession. Since some editorial style guides for the health professions, including the AMA's Manual for Authors and Editors, periodically

revise their editorial standards, examination programs must choose between maintaining an outmoded style or updating items and possibly affecting the existing item statistics. Health fields, in general, already utilize a highly technical vocabulary which, in itself, may preclude any effect from simple stylistic manipulation. Conversely, even minor editorial changes may exacerbate differences in performance.

Method

Subjects

The subjects were medical residents in a medical specialty who were taking a multiple-choice in-training examination. This examination provides diagnostic information to residents who intend to take a specialty board certification examination at a later date. The examination is timed, but most examinees are able to complete all test questions within the allotted period.

Test Materials

The items selected for the study were five-foil multiple choice questions chosen from the pretest sections of the in-training examination; these sections consist of questions that are being used for the first time to gather data on the item performance. These pretest questions are not tallied as part of the examinees' test scores. Twenty items were selected to cover three stylistic manipulations: the use of abbreviations (7 items), the use of numbers or Greek symbols (7 items), and the use of generic versus generic plus proprietary drug names (6 items). These particular stylistic manipulations were selected since they are encountered

most frequently in the examination program involved and were, therefore, of interest to content reviewers, editors, and psychometricians.

Each item that was selected was revised to create another item with briefer wording than the original item. Each item thus existed in a long version and a short version, and the matched pair differed only on the single stylistic variable under investigation for that pair. An attempt was made to select relatively common medical abbreviations, symbols, and drug names rather than esoteric ones in order to minimize the possibility of a "floor" effect resulting from difficult content. Therefore, the abbreviations in the study included phrases such as cerebrospinal fluid, blood urea nitrogen test, creatine phosphokinase, and deoxyribonucleic acid. Items with symbols employed Greek alphabet letters, numerals, degree and percent signs. Drug name items included common drugs such as furosemide, lidocaine, phenytoin, hydrochlorothiazide, and phenylephrine.

The content distribution of the twenty pairs of items corresponded proportionately to the content specifications outlined for the examination: there were four experimental items in each of five major content areas. The twenty pairs of items were also distributed across three cognitive taxonomy levels so that each of the five content areas contained an item under investigation in at least two of the three different cognitive levels. The twenty pairs of items also included two pairs of items requiring x-rays and graphs, roughly the same percentage as the overall examination.

A random replications design was planned. The 40 items under

investigation (20 long versions and 20 short versions) were divided into four sets: two 10-item sets of long versions and two 10-item sets of short versions. Each replication compared one long version with its matched short version. Each candidate responded to only one set of ten items presented in either the long or the short version. These items were inserted throughout the pretest sections and were scattered as much as possible to avoid systematic fatigue or practice effects.

Procedures

The examination was administered by program directors of training programs for the specialty involved. Test booklets were distributed in a spiraled sequence to the examinees so that each version of an item was answered by a random selection of one fourth of the candidates within each test center. The number of respondents to each set of experimental items ranged from 264-272. A random selection of examinee responses was dropped from some sets to ensure equal numbers of responses to each set. The examinees were not informed that pretest material was being incorporated into the examination or that multiple forms of questions were being tried out.

Item difficulty (p) and discrimination (biserial correlation) indices were calculated for each item with the total test score used as the criterion for the discrimination index. For each replication, a two-way analysis of variance (ANOVA) was performed on the sum of the correct responses to the ten experimental items. The first factor in the analysis was form (short, long) and the second

was ability (high, low). Ability was defined on the basis of the total number of scored items answered correctly, not including the pretest questions. Two different criteria were used for the ability factor: the first ANOVA used the top 50% and the bottom 50% of examinees, the second used the top 25% and the bottom 25%.

Individual item pairs were also analyzed by a two-way ANOVA using the 50% ability groups as the second factor. A Bonferroni test of significance was then applied to the differences across forms for each item pair. A similar two-way multivariate analysis of variance (MANOVA) was also performed with a vector of ten-item scores as the dependent variables. The MANOVA yielded essentially the same results as the ANOVA.

Results

The ANOVA using the 50% ability groups demonstrated significance of the model for replication 1 ($F = 23.31$, $p = .0001$) and replication 2 ($F = 8.88$, $p = .0001$) as shown in Table 1. However, this overall significance of the model in both replications is due to the expected significance of the ability factor as well as to the large number of candidates responding to each form. In examining the F test of the main effect due to form in replication 1, there appeared a form difference significant at the .03 level ($F = 4.57$). In this replication, the marginal mean for the ten items in the short form was 5.93 ($SD = 1.41$) and for the long form 5.62 ($SD = 1.67$). In replication 2, the mean on the short form ($\bar{X} = 6.01$, $SD = 2.80$) was exactly the same as that on the long form ($\bar{X} = 6.01$, $SD = 2.94$); thus, the zero mean square value. On neither replication was there a significant form/ability interaction.

Since a form/ability interaction might be masked by using the entire pool

of examinees, a different analysis was performed with the pool of examinees segmented on the basis of total scored item performance into the top 25% and the bottom 25%. Once again, no interaction effects were noted. In fact, in this analysis, the form difference that was significant in replication 1 did not prove significant ($F = 2.09$).

When the items were considered by type of stylistic manipulation, a similar lack of effect was noted. The three types of items did not perform differently on the different forms. As shown in Table 2, the abbreviation items and the drug items had similar average difficulty levels over both forms: .629 and .643, .625 and .621. The lower average difficulty in the symbol items (.496 and .529) is due to one especially difficult item pair. Each type of item had similar average discrimination levels on the different forms, with the largest difference (.294 and .244) appearing on the symbol items (as shown in Table 3).

The ANOVA comparison of individual pairs of items revealed no significant differences in performance across forms. The largest significance values are shown in Table 4. It should be noted that with the experimentwise error set at .05, these p values must be compared to a .005 (.05/10) significance level. Although these p values are not significant given the ten experimental items, symbol items seem to demonstrate more susceptibility to stylistic manipulations than the other two item types.

Discussion

This study was deliberately narrow in scope, covering only some typical kinds of modifications suggested by test editors. In addition, the number of variables is large compared to the small number of items used. However, since the number of in-training, licensure, and certification examinations in all

health fields has risen dramatically in recent years, the effect of these editorial changes on items is worth knowing.

The results present no evidence that these types of stylistic manipulations affect examinee performance on individual items. The performance of the three types of stylistic changes, by difficulty and by discrimination, also reveals no major difference in the long and short versions. These results could lead to a conclusion that health professionals have attained such a high degree of familiarity with the abbreviations, symbols, and drug names that these minor stylistic differences pose no problem for these examinees.

If test assemblers want to know that minor editing changes will not affect test statistics, of equal interest is whether the individual items can be used interchangeably when equating procedures and other sophisticated statistical analyses are performed. In this case, the results are mixed since there was statistical significance across forms in replication 1. Given the lack of any significant form effect in replication 2, the form effect in replication 1 could be due to the large sample size. This assumption seems reasonably justified in light of the nonsignificant ANOVA results obtained when only the top 25% and bottom 25% of examinees were used. However, further investigation would be required to establish under what circumstances form differences might appear.

While a sufficiently powerful test can detect small differences, the question becomes what is the practical significance of these results. Although we might postulate that test items with the most explicit information are least ambiguous and therefore fairest to all examinees in an examination setting, this may not necessarily be true. Perhaps both forms of the stylistic features used in this study are so commonplace and familiar that not

even the less able examinees are penalized by one form or the other. Abbreviations, in this case, may be considered part of the medical jargon learned in training. Of course, this result has content implications since not a few health professions explicitly test the knowledge of abbreviations in their examinations. The similar lack of difference in drug names may be more a testament to the changing practice in pharmacy where generic drugs are now routinely dispensed in many states. Even with the items using symbols, the short items are slightly (though not statistically significantly) easier, perhaps because the answer is more easily recognized in symbol form.

These results should interest editors, item writers, test developers, and researchers who have been concerned that minor stylistic changes will affect examinee performance on individual items. This kind of difference did not appear consistently in this study. Unfortunately, it is not possible to say definitively that there are no differences in the way examinees respond to items that are presented differently since form differences might occur with a different sample of items or examinees. Test assemblers still cannot assume that stylistic changes will never make a difference. Even if there is no difference on an item level, differences might exist on a test form level. And the chance that such differences might occur is enough to warrant caution where important decisions are being made about examinees.

References

- Board, C. and Whitney, D.R. (1972). The effect of selected poor item-writing practices on test difficulty reliability and validity. Journal of Educational Measurement, 9, 225-233.
- Bornstein, H. and Chamberlain, K. (1970). An investigation of the effects of "verbal load" in achievement tests. American Educational Research Journal, 7, 597-604.
- Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. Educational and Psychological Measurement, 44, 551-561.
- Haladyna, T.M. and Downing, S.M. (1985). A quantitative review of research on multiple-choice item writing. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
- Jerman, M.E. and Mirman, S. (1971). Linguistic and computational variables in problem solving in elementary mathematics. Educational Studies in Mathematics, 5, 317-362.
- McMorris, R.F., Brown, J.A., Snyder, G.W., and Pruzek, R.M. (1972). Effects of violating item construction principles. Journal of Educational Measurement, 9, 287-295.
- Millman, J. (1978). Determinants of item difficulty: A preliminary investigation. CSE Report No. 114, Center for the Study of Evaluation, UCLA, July.
- Plake, J. and Huntley, R.M. (1984). Can relevant grammatical clues result in invalid test items? Educational and Psychological Measurement, 44, 687-696.

Table 1
Summary Table*

	<u>MS</u>	<u>df</u>	<u>F</u>	<u>p value</u>
<u>Replication 1</u> (10 pairs of items)				
Model	64.95	3	23.31	.0001
Form (Short, Long)	12.73	1	4.57	.03
Ability (High, Low)	182.01	1	65.33	.0001
Form*Ability	0.12	1	0.04	NS**
Error	2.79	524		
<u>Replication 2</u> (10 pairs of items)				
Model	16.67	3	8.88	.0001
Form (Short, Long)	0.0	1	0.0	NS
Ability (High, Low)	47.28	1	25.18	.0001
Form*Ability	2.73	1	1.46	NS
Error	1.88	524		

* N = 528 per replication (264 on long version, 264 on short version)

** NS = Not significant at the .05 level

Table 2
Item Difficulty (p)

	Long Form		Short Form	
	<u>p</u>	<u>SD</u>	<u>p</u>	<u>SD</u>
<u>Abbreviations</u>	.629	.182	.643	.173
Replication 1	.532	.082	.592	.091
Replication 2	.703	.213	.682	.223
<u>Symbols</u>	.496	.236	.529	.219
Replication 1	.605	.153	.634	.101
Replication 2	.223	.177	.267	.233
<u>Drugs</u>	.625	.293	.621	.271
Replication 1	.498	.351	.491	.303
Replication 2	.688	.293	.687	.274

Table 3
Item Discrimination (BIS R)

	Long Form		Short Form	
	<u>p</u>	<u>SD</u>	<u>p</u>	<u>SD</u>
<u>Abbreviations</u>	.327	.131	.323	.068
Replication 1	.400	.115	.310	.078
Replication 2	.273	.127	.333	.069
<u>Symbols</u>	.294	.071	.244	.094
Replication 1	.306	.059	.278	.058
Replication 2	.265	.120	.160	.461
<u>Drugs</u>	.272	.461	.285	.065
Replication 1	.310	.028	.220	.071
Replication 2	.253	.141	.318	.033

Table 4
Comparison of Item Pairs

	<u>Replication</u>	<u>$\bar{X}$ (Long)</u>	<u>$\bar{X}$ (Short)</u>	<u>p value*</u>
<u>Abbreviations</u>				
Pair 1	1	.500	.545	NS
2	1	.470	.534	NS
3	1	.625	.697	NS
4	2	.788	.784	NS
5	2	.716	.697	NS
6	2	.902	.879	NS
7	2	.405	.367	NS
<u>Symbols</u>				
Pair 1	1	.716	.636	.05
2	1	.693	.735	NS
3	1	.352	.470	.006
4	1	.697	.693	NS
5	1	.568	.636	NS
6	2	.348	.432	.05
7	2	.098	.102	NS
<u>Drugs</u>				
Pair 1	1	.250	.277	NS
2	1	.746	.705	NS
3	2	.254	.280	NS
4	2	.890	.883	NS
5	2	.777	.795	NS
6	2	.830	.788	NS

* These values should be compared to .005 (.05/10). By this criterion, none are significant.